Identification and conservation of novel Long Noncoding RNAs in cattle using RNASEq data

Kedra, D. 1, Bussotti, G. 1, Prieto, P. 1, Sørensen, P. 2, Bagnato, A. 3, Mckay, S.D. 4, Schnabel, R. 4, Taylor, J.F. 4, Guigo, R. 1, Notredame, C 1

1Centre de Regulacio Genomica (CRG), Spain, 2Aarhus University, Denmark, 3Università degli Studi di Milano, Italy, 4University of Missouri, USA,

This presentation represents the views of the Authors, not the EC. The EC is not liable for any use that may be made of the information.
IncRNA prediction strategies

- RNASeq based
- Homology based (Human Gencode 2 cow)
- Homology based (cow RNASeq filtered predictions vs several mammalian genomes)
Bovine data

• 30 Liver samples (@CRG) 2x 96bp, 2 500Mr
• 28 liver, small intestine, skeletal muscle samples 80bp x 1/2 (Jerry Taylor)
• 15 ovary samples 51bp x1 (Milano)
• 84 udder/muscle samples (Denmark)
• 1.2M bovine ESTs from NCBI (complementary to RNASeq)
Bos strategy 1

- mapping both with in house RNASeq mapping pipeline
  GRAPe
- GSNAP
- GMAP for ESTs (filtered by seqclean from PASA)
- cufflinks gene models
- merge models
- remove genes overlapping with known genes from ENSEMBL 72
Bos results merged cufflinks
Remove genes overlapping with ENSEMBL 72
Bos strategy 2

- remove 1 exon genes and transcripts shorter than 150bp
- extract transcript sequences
- cluster transcripts at 90% identity (usearch)

RESULT 1:
- 15,356 transcripts from 9,775 genes
- check for repeats, ORFs, sizes etc.
- find putative non-coding transcripts
Sanity check

- blastn all transcripts against human GENCODE 17 lncRNAs
- top hits (e= 0.0) include MALAT1, KCNQ1, KLHL7, MMP24 etc.
Repeat Masking

- 40% of all our transcript sequence was masked by RepeatMasker (*)
- Selected transcripts with no more 20% of the repetitive sequence
  - RESULT: 4541 transcripts from 3255 genes
  - Sanity check: still got MALAT1 etc.

- * Gencode 17 human: 25% repetitive
Homology prediction: PipeR

Mapping overview
Human Gencode 2 cow: homology

- Take all Gencode 17 IncRNA transcripts
- Run PipeR using cow genome as target

RESULT:
- 4210 human transcripts have 4758 homologues from 3195 genes in cow
Results comparison: geneid coding potential
Homology vs RNASeq

- Little overlap:
  - 3195 genes have just 295(!) overlaps with strict non-repetitive RNASeq transcripts
  - But 1423 of them have overlap with all cufflinks 2 exons or more RNASeq based transcripts (at 10% or more)

- Probably in both approaches we do not get complete gene models but just gene fragments
Cow IncRNA (RNASeq) vs other species

- So far 874 cow queries queries produced:
  - 384 human
  - 404 pig
  - 211 mouse
- Give it another week or so...
Thank you for your attention!

And everybody involved during the study!

- Centre de Regulacio Genomica (CRG), Spain,
  - Bussotti, G., Prieto, P. Guigo, R., Notredame, C

- Aarhus University, Denmark,
  - Sørensen, P.

- Università degli Studi di Milano, Italy
  - Bagnato, A.

- University of Missouri, USA,
  - Mckay, S.D., Schnabel, R., Taylor, J.F.